

How Good a Map? Putting Small Area Estimation to the Test

Gabriel Demombynes, Chris Elbers, Jean O. Lanjouw, and Peter Lanjouw¹

Abstract

This paper examines the performance of small area welfare estimation. The method combines census and survey data to produce spatially disaggregated poverty and inequality estimates. To test the method, predicted welfare indicators for a set of target populations are compared with their true values. The target populations are constructed using actual data from a census of households in a set of rural Mexican communities. Estimates are examined along three criteria: accuracy of confidence intervals, bias and correlation with true values. We find that while point estimates are very stable, the precision of the estimates varies with alternative simulation methods. While the original Elbers et al (2002, 2003) approach of numerical gradient estimation yields standard errors that seem appropriate, some computationally less-intensive simulation procedures yield confidence intervals that are slightly too narrow. Precision of estimates is shown to diminish markedly if unobserved location effects at the village level are not well captured in underlying consumption models. With well specified models there is only slight evidence of bias, but we show that bias increases if underlying models fail to capture latent location effects. Correlations between estimated and true welfare at the local level are highest for mean expenditure and poverty measures and lower for inequality measures.

Keywords: Poverty, Inequality, Small Area Estimation

JEL Classification: C13, C88, D31, I32, O15, R13

World Bank Policy Research Working Paper 4155, March 2007

The Policy Research Working Paper Series disseminates the findings of work in progress to encourage the exchange of ideas about development issues. An objective of the series is to get the findings out quickly, even if the presentations are less than fully polished. The papers carry the names of the authors and should be cited accordingly. The findings, interpretations, and conclusions expressed in this paper are entirely those of the authors. They do not necessarily represent the view of the World Bank, its Executive Directors, or the countries they represent. Policy Research Working Papers are available online at <http://econ.worldbank.org>.

¹ World Bank, Free University of Amsterdam, UC Berkeley and World Bank. We are grateful to Martin Ravallion and Danny Pfeffermann for comments and suggestions. The views in this paper are the authors' and should not be interpreted to reflect those of the World Bank or affiliated institutions.

1 Introduction

This paper examines the performance of a method for producing small area estimates of the spatial description of economic welfare. The methodology is described in Elbers, Lanjouw and Lanjouw (2002, 2003), henceforth referred to as ELL (2002). These “poverty maps” offer the promise of generating useful data about poverty and inequality at the local level, information which has potential applications in both the policy and research spheres. In this paper, an unusual data set is used to compare community-level welfare measures estimated using the small area estimation method against measures created from direct observations of household expenditure collected over the entire population within those communities.

Poverty maps have two sets of uses. They can be used as tools for geographical targeting of social spending. In a number of countries they have been used by governments and non-governmental organizations to identify those areas where the poor are concentrated as a first step towards directing resources to the poor. While policymakers in wealthy nations are accustomed to having information about local level conditions and welfare readily at hand, in the typical less developed country, information compiled at the local level is scarce and only available through specialized surveys. In such environments poverty maps are a potentially valuable resource.

On the research front, poverty maps have a variety of applications. With the resurgent interest in economic growth theory, and in particular the focus on inequality’s role, spatial profiles of welfare within a country can be useful. Poverty maps can also be used to investigate the spatial relationship between poverty and a variety of outcomes, including health and crime. The research applications for poverty maps are particularly strong when poverty maps can be produced for multiple years in a single country. In such cases poverty maps can be employed for policy evaluation.

The method examined here has been employed for a number of countries, and the resulting poverty maps have been utilized by both policymakers and researchers.² The growing popularity of the methodology adds to the need for a validation exercise.

The analysis in this paper compares the *predicted* poverty and inequality rates produced by the methodology for groups of rural Mexican communities to the *actual* poverty and inequality rates in those communities. One strength of the small area estimation approach is that it produces confidence intervals for its estimated welfare measures. An important objective in this paper is to assess to what degree the confidence intervals produced by the ELL method capture the distribution of error in the point estimates. Bias in the point estimates is also examined. The paper is organized as follows. Section 2 details the poverty mapping methodology. Section 3 describes the data employed, Section 4 sketches the validation exercise, and Section 5 presents the results. Section 6 concludes with a discussion of results and their implications.

2. Methodology

This section reviews the poverty mapping methodology, which is explained in more detail in ELL (2002).³ The basic approach is straightforward and typically involves a household survey and a population census as data sources. First, the survey data are used to estimate a prediction model for either consumption or incomes. The selection of explanatory variables is restricted to those variables that can also be found in the census (or some other large dataset) or in a tertiary dataset that can be linked to both the census and survey. The parameter estimates are then applied to the census data, expenditures are predicted, and poverty (and other welfare) statistics are derived. The key assumption is that the models estimated from the survey data apply to census observations. The first stage begins with an association model of per capita household expenditure for a household h in location c , where the explanatory variables are a set of observable characteristics:

² Poverty Maps based on this method are now underway or completed in more than 30 developing countries. Early examples include Alderman et al. (2002), and Mistiaen, Ozler, Razafimanantena and Razafindravonona (2002). See also Demombynes et al (2002).

$$(1) \quad \ln y_{ch} = E[\ln y_{ch} | \mathbf{x}_{ch}] + u_{ch}.$$

The locations correspond to the survey clusters as they are defined in a typical two-stage sampling scheme. The observable characteristics must be found as variables in both the survey and the census or in a tertiary data source that can be linked to both data sets.⁴

Using a linear approximation to the conditional expectation, the household's logarithmic per capita expenditure is modeled as

$$(2) \quad \ln y_{ch} = \mathbf{x}'_{ch} \boldsymbol{\beta} + u_{ch}.$$

The vector of disturbances, \mathbf{u} , is distributed $F(0, \boldsymbol{\Sigma})$. The model in (2) is estimated by Generalized Least Squares using the household survey data. In order to estimate the GLS model, $\boldsymbol{\Sigma}$, the associated error variance-covariance matrix, is estimated. Individual disturbances are modeled as

$$(3) \quad u_{ch} = \eta_c + \varepsilon_{ch},$$

where η_c is a location component and ε_{ch} is a household component. This error structure allows for both spatial autocorrelation, i.e. a “location effect” for households in the same area to the extent that it is not already covered by location-level explanatory variables, and heteroskedasticity in the household component of the disturbance. The two components are uncorrelated and (by construction) uncorrelated with observable characteristics in the regression equation.

The model in (2) is first estimated by simple OLS. The residuals from this regression serve as estimates of overall disturbances, given by \hat{u}_{ch} . These residuals are decomposed into uncorrelated household and location components:

³ Early variants of the methodology were presented in Hentschel et al (2000) and Elbers, Lanjouw and Lanjouw (2000). These earlier versions differ in important ways with the approach outlined in ELL(2002).

$$(4) \quad \hat{u}_{ch} = \hat{\eta}_c + e_{ch}.$$

The estimated location components, given by $\hat{\eta}_c$, are the within-cluster means of the overall residuals. The household component estimates, e_{ch} , are the overall residuals net of location components. Additional parameters are estimated: $\hat{\sigma}_\eta^2$, the variance of η_c and $\hat{V}(\sigma_\eta^2)$, the variance of σ_η^2 .⁵

To allow for heteroskedasticity in the household component, a logistic model of the variance of ε_{ch} conditional on a set of variables, \mathbf{z}_{ch} , is estimated, bounding the prediction between zero and a maximum, A , set equal to $(1.05) * \max\{e_{ch}^2\}$:

$$(5) \quad \ln\left[\frac{e_{ch}^2}{A - e_{ch}^2}\right] = \mathbf{z}_{ch}^T \hat{\alpha} + s_{ch}.$$

Letting $\exp\{\mathbf{z}_{ch}^T \hat{\alpha}\} = B$ and using the delta method, the model implies a household specific variance estimator for ε_{ch} of

$$(6) \quad \hat{\sigma}_{\varepsilon, ch}^2 \approx \left[\frac{AB}{1+B}\right] + \frac{1}{2} \text{Var}(s) \left[\frac{AB(1-B)}{(1+B)^3}\right].$$

This heteroskedasticity model generates a vector of coefficient estimates, $\hat{\alpha}$, and the variance-covariance matrix, $\hat{V}(\hat{\alpha})$. The coefficient estimates are used to predict $\hat{\sigma}_{\varepsilon, ch}^2$, the household-specific term for the variance of ε_{ch} .

These error calculations are used to produce two square matrices of dimension n , where n is the number of survey households. The first is a block matrix, where each block corresponds to a cluster, and the cell entries within each block are $\hat{\sigma}_\eta^2$. The second

⁴ Note that these variables need not be exogenous.

⁵ See Appendix 1 of Elbers et al (2002) for details.

is a diagonal matrix, with household-specific entries given by $\hat{\sigma}_{\varepsilon, ch}^2$. The sum of these two matrices is $\hat{\Sigma}$, the estimated variance-covariance matrix for the original model given by equation (2). Once this matrix has been calculated, the original model is estimated by GLS.

In the second stage predicted log expenditures and subsequently local-level estimates of poverty and their accompanying standard errors can be generated via several routes. Elbers et al (2002) describe a method based on numerical gradient estimation. An alternative approach known as parametric bootstrapping (Pfeffermann and Tiller, 2005) has been found to yield closely similar results and proceeds as follows.⁶ A series of simulations are conducted, where for each simulation r a set of first stage parameters are drawn from their corresponding distributions estimated in the first stage. A set of beta and alpha coefficients, $\tilde{\beta}^r$ and $\tilde{\alpha}^r$, are drawn from the multivariate normal distributions described by the first stage point estimates and their associated variance-covariance matrices. Additionally, $(\tilde{\sigma}_{\eta}^2)^r$, a simulated value of the variance of the location error component is drawn.⁷ Combining the alpha coefficients with census data, for each census household $(\tilde{\sigma}_{\varepsilon, ch}^2)^r$, the household-specific variance of the household error component, is estimated. Then, for each household simulated disturbance terms, $\tilde{\eta}_c^r$ and $\tilde{\varepsilon}_{ch}^r$, are drawn from their corresponding distributions.⁸ A value of expenditure for each household, \hat{y}_{ch}^r , is simulated based on both predicted log expenditure, $\mathbf{x}'_{ch}\tilde{\beta}^r$, and the disturbance terms:

⁶ We will see below that while the methods yield very similar point estimates, the approach employed in ELL (2002) produces slightly wider (and possibly more plausible) confidence intervals. In Appendix 1 we outline yet a third approach that yields confidence intervals that also more closely track those obtained with the method outlined in ELL (2002).

⁷ The $(\tilde{\sigma}_{\eta}^2)^r$ value is drawn from a gamma distribution defined so as to have mean $\hat{\sigma}_{\eta}^2$ and variance $\hat{V}(\sigma_{\eta}^2)$.

⁸ Non-normality is allowed for in the distribution of both η_c and ε_{ch} . For example, for each distribution, a Student's t-distribution can be chosen with degrees of freedom such that its kurtosis most closely matches that of the first stage residual components, $\hat{\eta}_c$ or ε_{ch} . An alternative, semi-parametric, approach can also be adopted in which standardized residuals are drawn from the first-stage survey residuals.

$$(7) \quad \hat{y}_{ch}^r = \exp(\mathbf{x}_{ch}' \tilde{\boldsymbol{\beta}}^r + \tilde{\eta}_c^r + \tilde{\varepsilon}_{ch}^r).$$

Finally, the full set of simulated per capita expenditures, \hat{y}_{ch}^r , are used to calculate estimates of the welfare measures for each target population.⁹

This procedure is repeated R times drawing a new $\tilde{\boldsymbol{\alpha}}^r$, $\tilde{\boldsymbol{\beta}}^r$, $(\tilde{\sigma}_\eta^2)^r$ and disturbance terms for each simulation. For each subgroup, the mean and standard deviation of each welfare measure are calculated over all $r=1, \dots, R$ simulations. For any given location, these means constitute our point estimates of the welfare measure, while the standard deviations are the standard errors of these estimates.

There are two principal sources of error in the welfare measure estimates produced by this method.¹⁰ The first component, referred to as *model error* in ELL (2002), is due to the fact that the parameters from the first-stage model in equation (2) are estimated. The second component, termed *idiosyncratic error*, is associated with the disturbance term in the same model, which implies that households' actual expenditures deviate from their expected values. While population size in a location does not affect the model error, the idiosyncratic error increases as the number of households in a target subgroup decreases.

3. Data

The analysis in this paper uses data collected as part of the targeting and evaluation program of PROGRESA, a health, education, and nutrition program of the Mexican government. Assignment to PROGRESA for households in these communities was randomized by community; a census of all households in 506 communities was conducted in November 1997, 320 were integrated into PROGRESA in late spring of 1998, and three follow up surveys (complete censuses) of households in all 506 communities were conducted in 1998 and 1999. Additionally, a survey was conducted in

⁹ These calculations are performed using household size as weights, implicitly assuming that expenditure is distributed uniformly within households. The same methodology could be applied using equivalence scales to capture alternative intrahousehold distributional assumptions.

March 1998, before PROGRESA was introduced to treatment communities. The March survey included a fairly detailed expenditure survey.

This paper employs household characteristic data from the November 1997 survey and an expenditure aggregate constructed using the March 1998 survey.¹¹ While it would be possible to undertake the analysis using income data from the November survey, the expenditure data is preferred for two reasons. First, the income data is very noisy. A substantial fraction of households report no income at all, and the income data shows no correlation with the March expenditure aggregate. The March expenditure aggregate, in contrast, is highly correlated with an expenditure aggregate from the June 1999 survey (for control group households), suggesting that it is a fairly consistent measure of household welfare. Second, the applications of the ELL methodology thus far have most commonly used household expenditure or consumption as the basis for welfare analysis, following the consensus that given the potential for consumption smoothing, consumption is likely to be a better indicator of long-term welfare than income. While it would be preferable to have expenditure data collected at the same time as household characteristics data, the household variables used here are unlikely to change substantially over time. Consequently the time gap between the November and March surveys should not distort the analysis.

While detailed, the expenditure aggregate is less comprehensive than typical consumption aggregates developed from some surveys carried out in developing countries. It covers only cash expenditures and does not include figures for rent. The expenditure survey was not carried out in 14% of households interviewed in November 1997. These households, which are concentrated in a small number of communities, are not included in the analysis. The ten communities with fewer than 10 households with expenditure information are also not included, leaving 20544 households in 496 communities.

¹⁰ A third potential source of error is associated with computation methods. Elbers et al (2002) show that this can be set arbitrarily small by selecting a sufficiently large number of simulations.

¹¹ Most questions in the November 1997 survey were similar to those in the 2000 national Mexican census. They concerned household characteristics and recent income of the household.

4. Analysis

The approach used for the validation exercise is to estimate a first-stage model using a “pseudo-survey” drawn from the PROGRESA households, using a two-stage sampling procedure. Welfare measures are then predicted with target populations composed of groups of PROGRESA households. The PROGRESA communities themselves have too few households to produce meaningful confidence intervals for the estimates using the methodology. Previous experience, e.g. ELL (2002), has shown that standard errors are very large for target populations with less than a few hundred households. In order to generate a group of more suitably sized target populations, the communities were grouped at random into 20 target populations. Both the pseudo-survey and the target populations were drawn repeatedly, in order to generate estimates for a large number of target populations.

Specifically, the steps in the analysis were as follows:

- 1) A random sample of 50 localities was drawn from the 496 localities, with probability of selection proportional to the size of the locality. From each of the 50 localities, 10 households were selected at random. The data from these households (a total of 500) serve as a pseudo-survey.
- 2) The first-stage methodology described above was applied using the pseudo-survey. A set of explanatory variables for log per capita expenditure was selected from a candidate list. An additional set of explanatory variables which best explained estimated location effects were selected from a set of community-level averages.¹²
- 3) The 496 localities were grouped into 4 groups of 24 communities and 16 groups of 25 communities. These serve as the 20 target populations for the poverty mapping

¹² From equation (2) and (3) it is clear that the variance of the location effect η_c must be small if acceptable standard errors on welfare predictions are to be obtained. We have found that the inclusion of means of explanatory variables, calculated from the census for the relevant enumeration areas, reduces σ_η^2 considerably. See ELL(2002) for details and see also below.

analysis, and the location effect is modeled at the level of the localities. The target populations each cover an average of 1042 households.

- 4) True poverty and inequality rates were calculated for the 20 target populations based on actual per capita expenditure.¹³
- 5) The poverty mapping methodology was applied to predict poverty and inequality rates for the 20 target populations, using first-stage models estimated with the pseudo-survey.
- 6) The entire procedure was repeated 10 times, drawing a new pseudo-survey for each round of analysis.

The output of this procedure is a set of poverty and inequality estimates and associated standard errors for 200 target populations. To examine the sensitivity of the estimates to the error specification, two different specifications are used for the second-stage analysis. In the first, both the location component and the household component of the error are modeled as Student's t-distributions. For the second specification, a semi-parametric approach is used for both the location and the household components. In this semi-parametric approach, instead of drawing from a t-distribution, the standardized residuals are drawn from the first-stage survey residuals. For both specifications, the household component of the error is modeled as heteroskedastic, with the predicted log per capita expenditure as the sole explanatory variables.¹⁴

¹³ The poverty line was set to 159 pesos, the per capita expenditure of the median household in the full set of households. This corresponds roughly to PROGRESA's poverty-classification scheme; using discriminant analysis techniques based on household income, approximately 50% of households were initially classified as "poor" and thus qualified for PROGRESA.

¹⁴ Note that for the semi-parametric approach, it is the *standardized* residuals that are drawn from the observed distributions in the survey. These standardized residuals, with mean zero and variance equal one, are drawn and multiplied by the square root of the relevant simulated variance (of the location or household effect) to produce simulated residual values.

5. Results

First-Stage Results

OLS Regression results from the first-stage models are given in Appendix 2 Tables A1-A10. Across the ten pseudo-surveys used here, the R^2 ranges from 0.415-0.53 (see Table 1). The explanatory power of the models in this analysis is in the general range of models from past applications. The R^2 for models for particular strata ranged from 0.45 to 0.77 in Ecuador (Hentschel et al, 2000), 0.29 to 0.63 in Madagascar (Mistiaen et al, 2002), and 0.47 to 0.72 in South Africa (Alderman et al, 2002). The explanatory power achieved with the PROGRESA models is rather good given that the households in the PROGRESA communities are more homogenous than those within a stratum in a typical application. All the communities in the PROGRESA sample were selected for the program because they were poor and rural, based on indicators in the 1990 and 1995 censuses. Consequently, the households are more similar to one than another than the households in an entire stratum of a country.

Household size was used in all models, and some variables were selected in models for several pseudo-surveys, but there was generally little consistency in models chosen across pseudo-surveys. The estimated location effects were generally small, with variances ranging from 0.9% to 3.1% of the overall variance of the disturbance term after the addition of cluster-level means. This can also be seen in that the models achieved levels of explanatory power very close to what would be achievable with models that employed, instead, a cluster-level fixed-effects specification (see Table 1).

Second-Stage Results

5.1 Point Estimates and Precision

Tables 2 and 3 present illustrative results for the headcount rate based on two pseudo surveys: 2 and 3.¹⁵ These tables present for each of the 20 target populations a measure of the true headcount rate as well as the estimated headcount rate based on a variety of procedures. Column 1 presents estimates and standard errors based on the

¹⁵ These two pseudo-surveys have been chosen arbitrarily in order to avoid unnecessary repetition. Qualitative conclusions are unchanged if other, or all, pseudo-surveys are examined.

numerical gradient simulation procedure sketched out in Elbers et al (2002). Columns 2-4 present estimates based on the “parametric bootstrapping” (Pfeffermann and Tiller, 2005) procedure outlined in section 3 and are computed using the POVMAP2 software that has been purpose-written by Qinghua Zhao in the Research Department of the World Bank.¹⁶ The parametric bootstrapping results vary depending on whether disturbances are drawn from the empirical distribution (Column 2) or from parametric distributions (Column 3). The estimates in column 4 are based on a program written in SAS, based also on application of the procedure outlined in section 3 (with disturbances drawn from a parametric distribution), and are presented to illustrate that simulation based results do vary depending on different random number generating algorithms as well as seeds. Finally the results presented in Column 5 are based on an alternative, non-parametric, scheme outlined in Appendix 1.¹⁷

Point estimates differ only slightly across different simulation approaches. In Table 2, while the true headcount rate for target population 1 is 60.5% the estimated rate for this target population varies between 60.9% and 61.6% across the different estimation approaches. The approaches are more clearly at odds in terms of the estimated standard errors. In particular, standard errors deriving from the “parametric bootstrapping” procedure described in Section 3 and summarized in Columns 2-4, tend to be somewhat smaller than those based on the numerical gradient method described in ELL(2002) – Column 1 - and the non-parametric approach of Appendix 1 (Column 5). In the case of pseudosurvey 2 the distinction is not of great significance: irrespective of methodology, the 95% confidence interval around each target population’s estimated headcount rate encompasses the true poverty rate in 19 out of 20 cases. However, with other pseudo surveys the distinction does matter. In Table 3, results are presented based on a model of consumption estimated from pseudosurvey 3. With this survey, the “classical” approach (Elbers et al, 2002) and the alternative approach outlined in the appendix yield three

¹⁶ POVMAP2 can be freely downloaded at <http://iresearch.worldbank.org>.

¹⁷ Note that these estimates do not show significant differences in poverty between target populations. This reflects both the relative homogeneity of the group of PROGRESA households, the random composition of target populations, and the small sizes of the target populations, about 1000 households. On the other hand, discriminating between poverty of the target populations is not the subject of the current paper and all standard errors are about the same size as one would get from survey-based estimates at the aggregate level..

cases where a target population’s true poverty rate falls outside the 95% confidence interval around the estimated poverty rate. But with the parametric bootstrap approach underpinning estimates in Columns 2-4 the failure rate is higher (7 cases). For this pseudosurvey the parametric bootstrapping approach appears to produce standard errors that are too “optimistic” - suggesting greater precision of estimates than is warranted.

Given this evidence of a tendency for the parametric bootstrapping procedure to produce confidence intervals that are somewhat too narrow, we employ from now on, unless noted explicitly otherwise, the non-parametric approach outlined in Appendix 1. Additional comparisons, not reported here, confirm that conclusions derived with this simulation procedure hold also for estimates based on the considerably more computationally-intensive numerical gradient approach outlined in ELL(2002). The important point to take away here is that simulation methods do seem to matter (with respect to standard errors, if not point estimates). Further research is underway to understand better why the different simulation methods do not always agree.¹⁸

Table 4 looks more closely at the confidence intervals estimated around welfare estimates produced with our non-parametric simulation scheme. If the confidence intervals accurately reflect the true uncertainty in the estimates, the fraction of cases of the “truth” falling within a confidence interval around an estimate should be approximately equal to the corresponding confidence level. Note however that twenty ‘target populations’ are drawn for each of the ten ‘surveys’ and so the experiments are not entirely independent.

For each welfare measure and each of the ten pseudosurveys the number of instances is counted when true welfare in each of the 20 target populations falls within two standard deviations around the target population’s estimated welfare level. For example, in the case of pseudo survey 1, the true welfare estimate (mean, headcount, squared poverty gap, and General Entropy Class inequality measure with parameter 0) always fall within the confidence interval around the estimated welfare measure. In Table 2 we saw that for pseudosurvey 2 this occurred 95% of the time (19 out of 20 cases) for the headcount, and Table 4 shows the same was observed for the mean, while

¹⁸ The most recent version of POVMAP2 now offers the user the choice of the “classical” numerical gradient or the parametric bootstrapping procedures outlined in Section 3. .

for the squared poverty gap and inequality calculated on the basis of the GE0 the truth always falls within the confidence intervals calculated around the estimates. On average, across all pseudo surveys the success rate is just under 95% for the mean consumption, headcount, and squared poverty gap measures, and just below 90% for the GE0 measure.

In Table 5 we consider how sensitive are our estimated standard errors to the presence of unobserved location effects. We saw in Table 1 that our preferred specifications for the different pseudosurveys were quite successful in proxying unobserved location effects ($\hat{\sigma}_\eta^2 / \hat{\sigma}_u^2$ ranges between 0.9% and 3%). How much larger would standard errors be if our underlying models had not been so successful in this respect? Table 5 compares estimates and standard errors on small area estimates of the headcount rate from pseudosurvey 2 based on two models: one with our preferred specification; and the other with a specification in which no census-mean variables were included.¹⁹ In the latter model the share of the variance of overall disturbance term that is attributable to the variance of the cluster component is now 11.9%, a four-fold increase over the 2.7% in the preferred model (Table 5). At the all-census level, the two models predict headcount rates of 61.9% and 61.5%, respectively, both virtually indistinguishable from the 61.1% actual headcount rate in the population. However, the standard error on the model with no location variables is now 0.024, up by more than two fifths from the standard error of 0.017 obtained with the preferred model. Part of the increase in the standard error is due to the fact that the explanatory power of the model with no location variables is lower than that of the preferred model. As a result, idiosyncratic error would be expected to be higher – see Section 2 and ELL(2002). However, at the level of the total population most of the idiosyncratic error will have cancelled out (poverty is being estimated over a population of more than 20,000 households). Thus the increase in the standard error from 0.017 to 0.024 is likely due mainly to the consequence of our failure to adequately capture unobserved location effects. At the target population level, standard errors are higher than at the level of the total population, irrespective of underlying models. Moving from the preferred specification to the model with no location variables, standard errors rise considerably,

¹⁹ Our calculations here are based on the numerical gradient “classical” simulation procedure.

and in some cases the percentage change is even greater than at the level of the total population. For example, standard errors across the two models rise by as much as 43% for target population 2 ($0.030 \times 1.43 = 0.043$). However, here, the changes in standard errors are reflecting both the influence of idiosyncratic error and our failure to capture location effects.

5.2 The Level of Location Effects

Note that the location effect η_c may include group effects at levels higher than the survey cluster. To see this consider the following model with group random effects at a ‘district’ level (v), as well as the cluster level (c):

$$\ln y_{vch} = \mathbf{x}'_{vch} \beta + \omega_v + \eta_{vc} + \varepsilon_{vch}$$

As before, the error components are uncorrelated. If clusters are the primary sampling unit, a district is sampled only indirectly, *viz.* if one of the sampled clusters happens to be located in that district. In a typical living standards survey there will only rarely be districts that have been sampled more than once in this way, making it impossible to separate the location effect in the sample into a ‘district effect’ ω and a ‘cluster effect’ η . Assume accordingly that a district is sampled at most once, and write $v(c)$ for the unique district sampled along with the cluster. The model now becomes

$$\ln y_{v(c)ch} = \mathbf{x}'_{v(c)ch} \beta + \omega_{v(c)} + \eta_{v(c)c} + \varepsilon_{v(c)ch}.$$

Or, with obvious relabelling:

$$\ln y_{ch} = \mathbf{x}'_{ch} \beta + \eta_c^* + \varepsilon_{ch},$$

where $\eta_c^* = \omega_{v(c)} + \eta_{v(c)c}$. Consequently, the estimated variance of the location effect in a model with only cluster-level random effects is in fact an estimate of $\sigma_\omega^2 + \sigma_\eta^2$, the combined group effects operating at the sample’s cluster level.

In the simulation phase the analyst has to choose whether the location effect estimated from the pseudosurvey should be applied at the cluster or the ‘district’ level. When there is no way of separating the location effect into a cluster and ‘district’ effect the best that one can do is to assume either that the effect is entirely a cluster-level effect, or that it occurs entirely at the district-level. The latter will be quite a conservative assumption as it will rule out that any part of the estimated location effect applies only at the cluster level. This approach might be considered as yielding an “upper-bound” on the standard error. The former will be “optimistic” in the sense that it will yield standard errors that could be under-estimates of the true-standard error – particularly if the location effect is big. In our setting, it does not make sense to apply the location effect at a level higher than the cluster, as the latter correspond to villages and these have been assembled *randomly* into 20 target populations. ELL (2002) illustrate in the more plausible setting of rural Ecuador, however, that when it is assumed that the location effect estimated at the cluster level applies entirely at a higher level (in Ecuador, at the *parroquia* level), then the idiosyncratic component of the standard error does rise appreciably. However, they also show that the impact on overall standard errors is negligible because in their setting – as in the present study – the size of the estimated location effect is small. If the introduction of cluster-means or other cluster-level variables is not successful in capturing group effects then the choice of level of aggregation at which to apply the location effect in the simulations can affect final results more substantially. In such a case there would be a larger range between the “optimistic” standard errors and the upper-bound estimates obtained by assuming that the location effect occurs entirely at the ‘district’ level.

5.3 Bias

Another way in which to gauge the reliability of small-area estimates of welfare is to consider whether there is evidence of bias - a systematic tendency for estimates to deviate from the truth in any way. Figures 1-4 show, for each target population and for four different welfare measures, the relationship between true welfare and the difference between true and estimated welfare. In Figure 1 we can see that there is some tendency for the estimation procedure to overestimate mean per-capita consumption for those

target populations with a true mean consumption level that is low, and to underestimate the mean consumption level of rich target populations. To see this note that when true consumption is low, the bias - defined here as “truth” minus estimated consumption, is negative – while it is positive when true average consumption is high. However this relationship is not strong. Overall, the average difference between the estimated mean consumption and true consumption is about 1.5 pesos: about 1% of the mean consumption level of the poorest target population. The bias is similarly modest for the headcount (Figure 2), squared poverty gap (Figure 3) and mean log deviation (General Entropy class measure with parameter 0) inequality measure (Figure 4).

The extent of bias in these estimates is related to the degree to which the model specification fails to capture location effects on the basis of census-mean variables or other variables intended to capture locality-level characteristics. As we saw in the preceding section and in Table 1, our model specifications are quite successful in removing the effect of latent community level characteristics, and as a result the bias in our estimates is quite modest. If we produce estimates that omit village-level census means, then the bias is accentuated. Figure 5 illustrates how the slope of the line capturing the extent to which headcount is overestimated in truly non-poor communities and the headcount is underestimated in truly poor communities becomes steeper when estimates are based on a consumption model that fails to capture unobserved location effects. The intuition behind this bias is quite straightforward: if there is a sizeable location effect, and our model fails to capture it, then there will be a tendency for poverty to be over-estimated in communities that are relatively well-off, given the explanatory variables in the model, i.e. that have large positive location effects. Part of the reason that the communities are well-off is likely attributable to community-wide characteristics of the community, and this will not be reflected in estimates based on a model that fails to capture the effect of those characteristics. As a result estimates will tend to overstate poverty of such communities. Conversely, in truly poor communities, part of the reason they are so poor will be due to the broader characteristics of the community. Again, if the consumption model does not capture the impact of those broader characteristics, there will be a tendency for estimated poverty to be an understatement of true poverty in the community. We see, therefore, that not only is there a strong incentive to proxy location

characteristics in order to improve the precision of estimates (Section 5.1), but also in order to minimize a systematic tendency to overstate poverty in truly non-poor communities and understate poverty in truly poor communities.

5.4 Correlation

A further way to consider the reliability of the small area estimates is to examine the correlation between the predictions and the true values. Table 6 shows simple pearson and spearman rank correlations between true and predicted values. Each cell shows the correlation between predicted welfare and true welfare across the 20 target populations. Rows represent alternative pseudosurveys and columns indicate alternative welfare measures. Correlations (both pearson and rank) are positive and reasonably high for mean consumption and the two poverty measures (headcount rate and squared poverty gap). In the case of inequality the correlations are much lower – presumably because the target populations vary very little in terms of true inequality. Indeed, households in the PROGRESA communities are more homogeneous than those within a stratum in a typical poverty mapping application. All the communities in the PROGRESA sample were selected for the program because they were poor and rural, based on indicators in the 1990 and 1995 censuses. Consequently, the households are more similar to one another than the households in an entire stratum of a country. This high level of homogeneity across households (and target populations) is a somewhat unusual feature of this empirical application. However, it might be expected to present a particularly difficult setting in which to implement the small-area estimation methodology and therefore does provide a useful (conservative) setting in which to gauge the methodology’s performance.

6. Discussion

The results presented here offer a rough test of the ELL methodology and point to some tentative conclusions that may inform future applications of the ELL welfare mapping method. In terms of the predictive power of the method, the results provide strong evidence that ELL estimates have important information content. Bias is low, the correlations between actual and predicted values of poverty indices and the mean are

generally positive and not insubstantial. For inequality figures, the results are generally weaker. Because the signal-to-noise ratio is lower in these inequality estimates, it is particularly important to take into account error in the estimates when applying them to research or policy applications.

The ability to provide confidence intervals is a crucial advantage to the ELL method as compared with alternative approaches to welfare mapping. In the analysis presented here, it was found that alternative simulation methods do influence the size of the estimated standard errors on welfare estimates. The numerical gradient approach, originally proposed in ELL(2002) was found to produce satisfactory standard errors, and similarly for the non-parametric simulation procedure outlined in Appendix 1. However, the parametric bootstrapping procedure described in Section 3 was found to yield standard errors that are somewhat understated. It is not entirely clear why this latter procedure should suffer from this propensity, and further research is needed to resolve this concern.

An important objective of this analysis has been to document how important it is, when applying small-area estimation methods, to think hard about possible unobserved, community-level, factors that may influence welfare outcomes. Experience with “poverty mapping” in a large number of countries indicates that inclusion of census-means as regressors in the underlying consumption model (and/or the inclusion of household variables that capture “network” effects, or of additional community-level variables from tertiary datasets such as administrative and GIS data) can go a long way towards helping to secure specifications in which unobserved location effects are kept small. The analysis here has shown that failure to capture such location effects in this way can lead to markedly higher standard errors and also an increase in bias.

It is important to recognize the limitations of the analysis in this paper. The data used here are less well-suited to poverty mapping than those usually employed. First, the expenditure aggregate used is less comprehensive than that found in a typical developing country survey, and the general quality of the data may be worse than, for example, data collected in a World Bank LSMS survey. This reduces the potential for variation in expenditure to be explained by observed variables. Second, the data all come from poor households in rural Mexico. Consequently, there is relatively little variation in

expenditure across households, and a relatively large fraction of the variation is due to measurement error or short-term fluctuations and cannot be explained by observable characteristics.

The problem associated with the small range of expenditures is compounded in this exercise by the fact that it was necessary to construct target populations by randomly assembling groups of communities. This resulted in a narrow spread of welfare measure values across the target populations. The ELL method is likely to produce estimates with a higher signal-to-noise ratio when the underlying population has greater variation in consumption.

All in all, the analysis presented here suggests that the details of poverty mapping matter. But the evidence does also suggest that the small area estimation procedure can provide useful, and reliable, estimates of welfare at fine levels of aggregation that survey data themselves would not be able to accommodate.

References

- Alderman, Harold, Miriam Babita, Gabriel Demombynes, Nthabiseng Makhatha, and Berk Özler. "How Small Can You Go? Combining Census and Survey Data for Mapping Poverty in South Africa, 2002." *Journal of African Economies*, 11: 3.
- Demombynes, Gabriel, Chris Elbers, Jenny Lanjouw, Peter Lanjouw, Johan Mistiaen and Berk Özler. 2002. "Producing a Better Geographic Profile of Poverty: Methodology and Evidence from Three Developing Countries." WIDER Discussion Paper no. 2002/39, The United Nations.
- Elbers, Chris, Jean O. Lanjouw, and Peter Lanjouw (2000) "Welfare in Villages and Towns: Micro-Measurement of Poverty and Inequality", Tinbergen Institute Working Paper No. 2000-029/2, Amsterdam, Netherlands.
- Elbers, Chris, Jean O. Lanjouw, and Peter Lanjouw (2002) "Micro-Level Estimation of Welfare", Policy Research Working Paper No. 2911, The World Bank, October 2002.
- Elbers, Chris, Jean O. Lanjouw, and Peter Lanjouw. 2003. "Micro-level Estimation of Poverty and Inequality." *Econometrica* 71:1, pp. 355-364.
- Hentschel, J., Lanjouw, J.O., Lanjouw, P. and Poggi, J. (2000) "Combining Census and Survey Data to Study Spatial Dimensions of Poverty: A Case Study of Ecuador", *World Bank Economic Review* 14(1): 147-166.
- Mistiaen, Johan, Berk Özler, Tiaray Razafimanantena, and Jean Razafindravonona. 2002. "Putting Welfare on the Map in Madagascar" World Bank Africa Region Working Paper Series No. 34, The World Bank.
- Pfeffermann, D. and Tiller, R. (2005) 'Bootstrap Approximation to Prediction MSE for State-Space Models with Estimated Parameters', *Journal of Time Series Analysis*, 25(6), November, 893-916.

Table 1
Diagnostics for 10 Pseudosurvey Consumption Models

Pseudosurvey	Sample Size	No. of Clusters	\bar{R}^2	$\hat{\sigma}_\eta^2 / \hat{\sigma}_u^2$	$R^2 / R_{f.e.}^2$
1	500	50	0.4678	0.0291	0.927
2	500	50	0.4593	0.0270	0.912
3	500	50	0.5274	0.0247	0.927
4	500	50	0.4151	0.0019	0.901
5	500	50	0.5176	0.0195	0.961
6	500	50	0.4766	0.0259	0.920
7	500	50	0.4549	0.0263	0.971
8	500	50	0.4205	0.0241	0.910
9	500	50	0.4910	0.0088	0.945
10	500	50	0.4193	0.0310	0.874

TABLE 2: Pseudosurvey 2

Targetpop	Truth	(1) ‘Classical’ Procedure (Elbers et al 2002)		(2) PovMap2 (non-parametric)		(3) PovMap2 (parametric)		(4) SAS-based Program		(5) Alternative Procedure (see Appendix)	
		μ	s.e.	μ	s.e.	μ	s.e.	μ	s.e.	μ	s.e.
1	0.605	0.614	0.030	0.616	0.027	0.609	0.029	0.611	0.025	0.612	0.037
2	0.568	0.616	0.030	0.622	0.028	0.621	0.028	0.613	0.027	0.616	0.039
3	0.572	0.621	0.032	0.624	0.032	0.619	0.029	0.614	0.029	0.613	0.040
4	0.636	0.636	0.031	0.635	0.024	0.630	0.024	0.627	0.027	0.640	0.036
5	0.612	0.586	0.034	0.585	0.034	0.592	0.033	0.591	0.034	0.584	0.041
6	0.640	0.641	0.031	0.638	0.033	0.641	0.032	0.638	0.029	0.639	0.038
7	0.621	0.568	0.034	0.565	0.035	0.573	0.035	0.572	0.036	0.569	0.038
8	0.647	0.643	0.036	0.644	0.035	0.645	0.033	0.640	0.032	0.626	0.048
9	0.610	0.592	0.029	0.595	0.030	0.599	0.032	0.597	0.033	0.589	0.039
10	0.675	0.609	0.033	0.609	0.034	0.615	0.030	0.612	0.031	0.596	0.038
11	0.603	0.609	0.038	0.605	0.034	0.607	0.030	0.607	0.029	0.606	0.038
12	0.568	0.681	0.037	0.690	0.031	0.685	0.030	0.677	0.033	0.680	0.046
13	0.647	0.623	0.033	0.629	0.029	0.631	0.030	0.623	0.032	0.630	0.038
14	0.604	0.591	0.035	0.599	0.029	0.594	0.030	0.592	0.030	0.583	0.043
15	0.576	0.618	0.036	0.619	0.029	0.625	0.030	0.614	0.030	0.625	0.039
16	0.595	0.613	0.030	0.614	0.029	0.616	0.027	0.608	0.024	0.611	0.038
17	0.553	0.564	0.038	0.565	0.030	0.569	0.029	0.561	0.031	0.553	0.043
18	0.589	0.634	0.039	0.633	0.029	0.636	0.033	0.638	0.033	0.629	0.043
19	0.676	0.638	0.037	0.639	0.029	0.642	0.023	0.637	0.025	0.656	0.039
20	0.613	0.654	0.030	0.653	0.029	0.656	0.027	0.657	0.029	0.651	0.036
Cases of truth falling outside the 2 s.e. interval		1		1		1		1		1	

TABLE 3: Pseudosurvey 3

Targetpop	Truth	(1) ‘Classical’ Procedure	(2) PovMap2 (non-parametric)	(3) PovMap2 (parametric)	(4) SAS-based Program (parametric)	(5) Alternative Procedure (See Appendix)
-----------	-------	---------------------------------	------------------------------------	--------------------------------	--	--

		(Elbers et al 2002)									
		μ	s.e.	μ	s.e.	μ	s.e.	μ	s.e.	μ	s.e.
1	0.605	0.555	0.030	0.554	0.023	0.555	0.030	0.554	0.034	0.554	0.022
2	0.568	0.570	0.037	0.569	0.024	0.570	0.037	0.560	0.040	0.568	0.026
3	0.572	0.544	0.033	0.544	0.030	0.544	0.033	0.531	0.043	0.0541	0.030
4	0.636	0.554	0.034	0.551	0.029	0.554	0.034	0.548	0.043	0.554	0.024
5	0.612	0.576	0.032	0.582	0.028	0.576	0.032	0.562	0.040	0.580	0.028
6	0.640	0.591	0.033	0.587	0.027	0.591	0.033	0.581	0.040	0.591	0.026
7	0.621	0.571	0.033	0.575	0.028	0.571	0.033	0.566	0.039	0.573	0.026
8	0.647	0.629	0.036	0.629	0.032	0.629	0.036	0.619	0.040	0.632	0.029
9	0.610	0.554	0.034	0.556	0.024	0.554	0.034	0.554	0.038	0.558	0.023
10	0.675	0.595	0.033	0.600	0.026	0.595	0.033	0.574	0.043	0.594	0.025
11	0.603	0.584	0.038	0.586	0.025	0.584	0.038	0.587	0.037	0.586	0.027
12	0.568	0.562	0.034	0.561	0.027	0.562	0.034	0.556	0.043	0.563	0.028
13	0.647	0.567	0.040	0.568	0.027	0.567	0.040	0.568	0.040	0.567	0.025
14	0.604	0.527	0.030	0.525	0.025	0.527	0.030	0.531	0.039	0.523	0.022
15	0.576	0.548	0.030	0.545	0.022	0.548	0.030	0.549	0.037	0.545	0.025
16	0.595	0.589	0.026	0.589	0.026	0.589	0.026	0.593	0.040	0.588	0.025
17	0.553	0.492	0.033	0.495	0.022	0.492	0.033	0.487	0.030	0.497	0.025
18	0.589	0.548	0.040	0.549	0.024	0.548	0.040	0.546	0.042	0.547	0.025
19	0.676	0.649	0.031	0.651	0.025	0.649	0.031	0.641	0.033	0.651	0.024
20	0.613	0.652	0.040	0.653	0.025	0.652	0.040	0.632	0.039	0.652	0.027
Cases of truth falling outside the 2 s.e. interval		3		7		7		7		3	

Table 4: Relative Frequency of True Target Population Welfare Falling Within 95% Confidence Interval Around Estimated Welfare

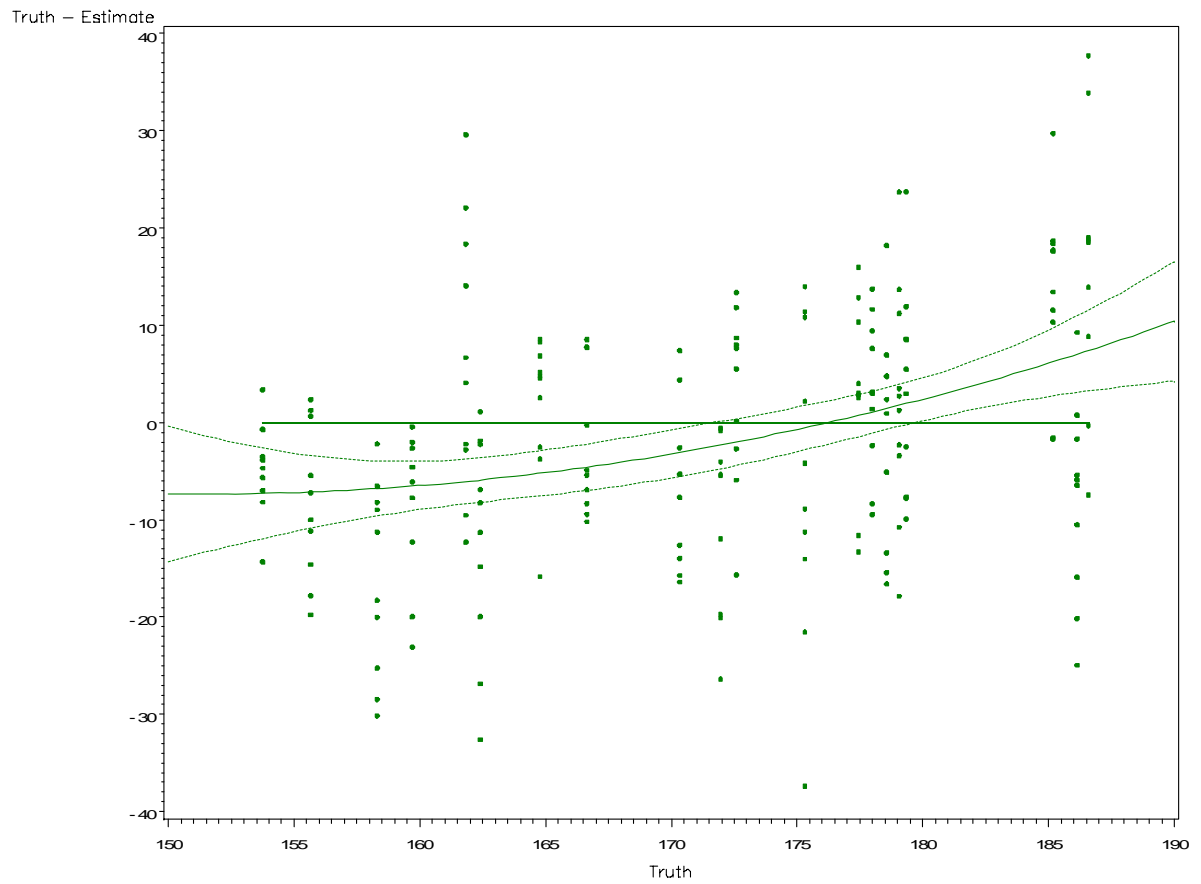
Survey	Mean	Headcount	FGT2	GE0
1	1.00	1.00	1.00	1.00
2	0.95	0.95	1.00	1.00
3	0.90	0.85	0.80	0.95
4	0.95	1.00	1.00	0.90
5	1.00	1.00	1.00	0.85
6	0.80	0.90	0.80	0.60
7	0.95	0.95	0.95	1.00
8	0.95	0.95	0.90	0.70
9	0.85	0.85	0.80	0.90
10	0.95	0.90	0.90	0.95
Overall	0.93	0.94	0.92	0.89

Table 5
Precision of Headcount Estimates with and without Location Variables
Numerical Gradient “Classical” Simulation
Pseudosurvey 2, POVMAP2 calculations

Village Code (sorted by true FGT0)	Population	True FGT0	I. Model with Location Variables		II. Model with no Location Variables		% change in standard error in moving from Model I. to Model II.
			Sample size=500 $\bar{R}^2 = 0.459$, $\sigma_\eta^2 / \sigma_u^2 = 0.027$		Sample size=500 $\bar{R}^2 = 0.413$, $\sigma_\eta^2 / \sigma_u^2 = 0.119$		
			Estimated FGT0	s.e.	Estimated FGT0	s.e.	
1	946	0.605	0.614	0.030	0.600	0.040	33%
2	1046	0.568	0.616	0.030	0.622	0.043	43%
3	1162	0.572	0.621	0.032	0.604	0.042	31%
4	991	0.636	0.636	0.031	0.598	0.041	32%
5	1061	0.612	0.586	0.034	0.609	0.042	24%
6	935	0.640	0.641	0.031	0.606	0.040	29%
7	932	0.621	0.568	0.034	0.602	0.046	35%
8	861	0.647	0.643	0.036	0.653	0.042	14%
9	871	0.610	0.592	0.029	0.615	0.038	31%
10	1219	0.675	0.609	0.033	0.622	0.040	21%
11	845	0.603	0.609	0.038	0.615	0.038	0%
12	992	0.568	0.681	0.037	0.624	0.044	9%
13	1289	0.647	0.623	0.033	0.623	0.039	18%
14	1271	0.604	0.591	0.035	0.624	0.045	29%
15	854	0.576	0.618	0.036	0.612	0.039	8%
16	1141	0.595	0.613	0.030	0.614	0.038	27%
17	1181	0.553	0.564	0.038	0.582	0.044	16%
18	820	0.589	0.634	0.039	0.616	0.045	15%
19	1060	0.676	0.638	0.037	0.623	0.038	3%
20	1008	0.613	0.654	0.030	0.637	0.040	33%
Total	20485	0.611	0.619	0.017	0.615	0.024	41%%

Figure 1: Checking for Bias

Difference between truth and estimated value of welfare estimate: MEAN
Ten different survey-based estimates for each of 20 target populations



Average difference: -1.49

Figure 2: Checking for Bias

Difference between truth and estimated value of welfare estimate: HEADCOUNT
Ten different survey-based estimates for each of 20 target populations

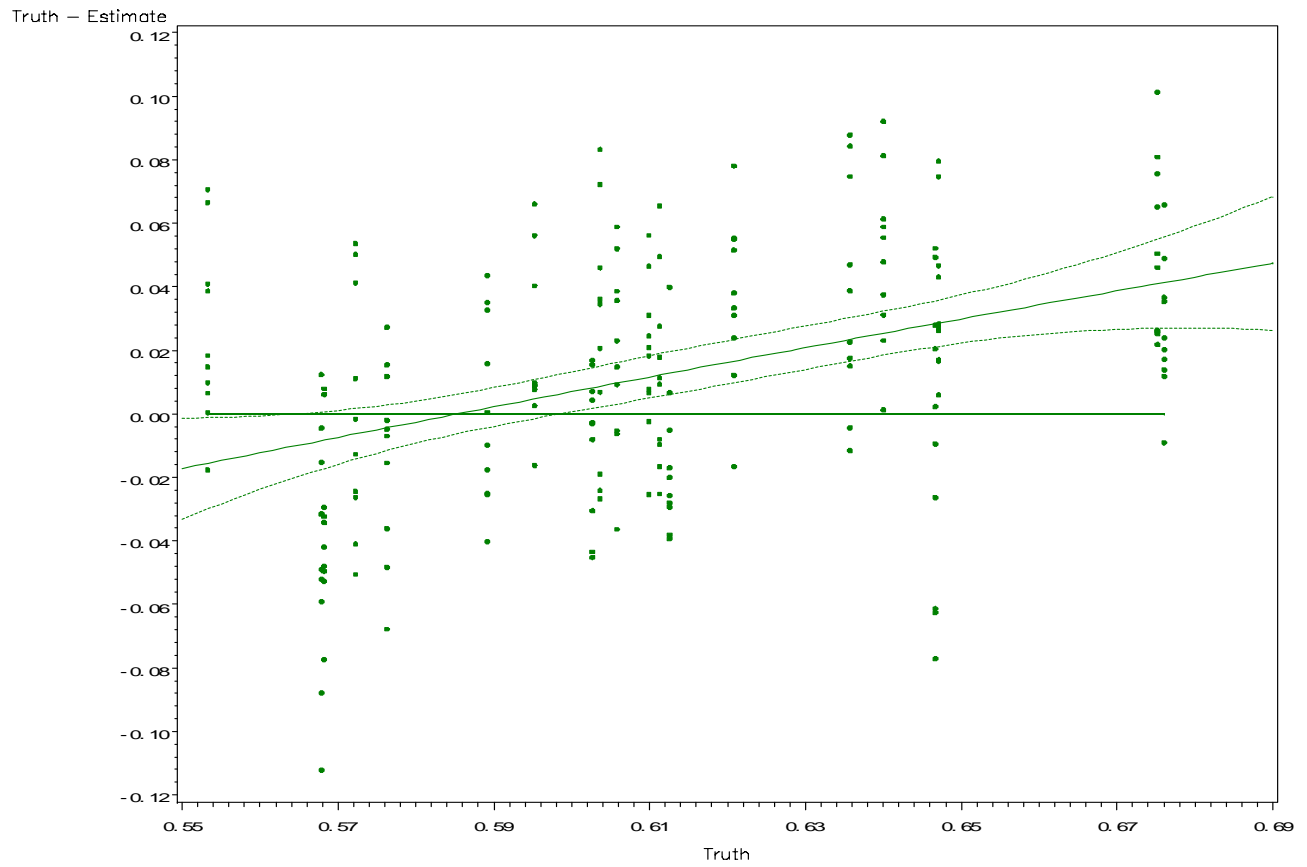
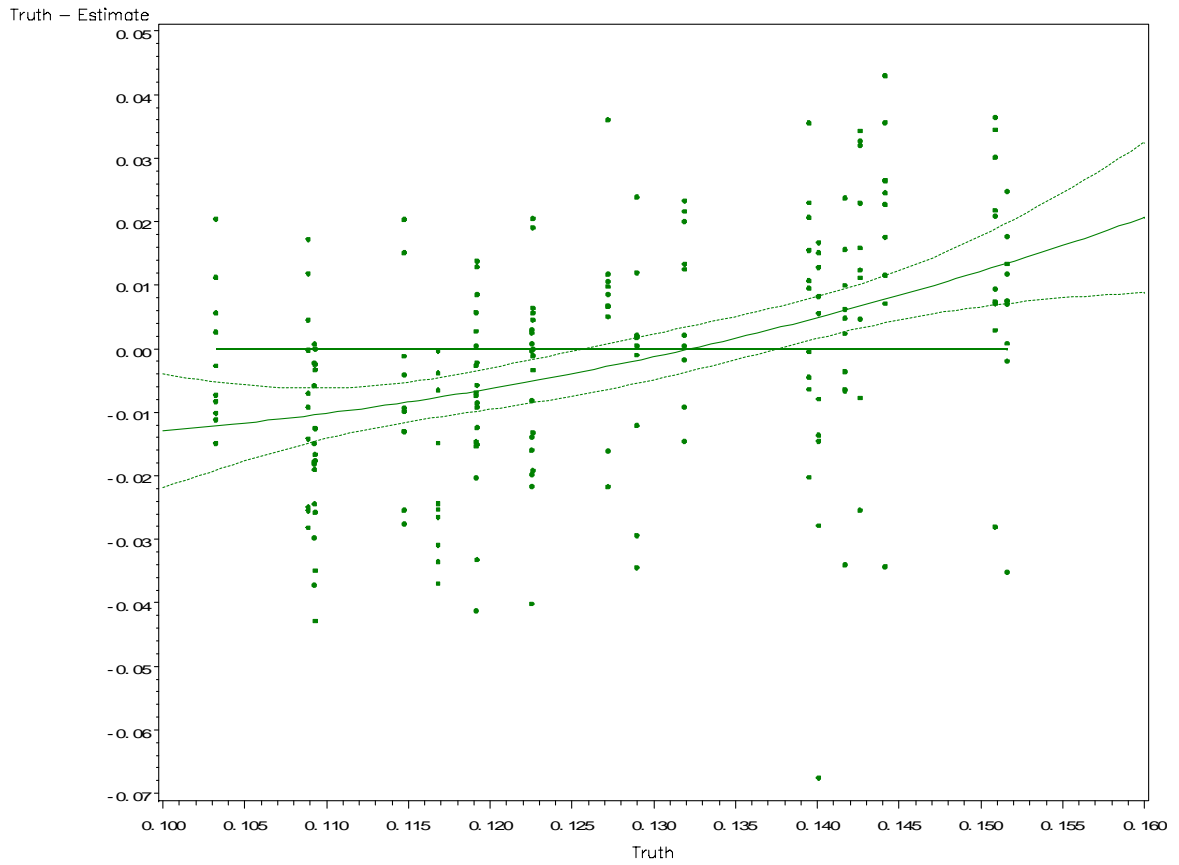


Figure 3: Checking for Bias

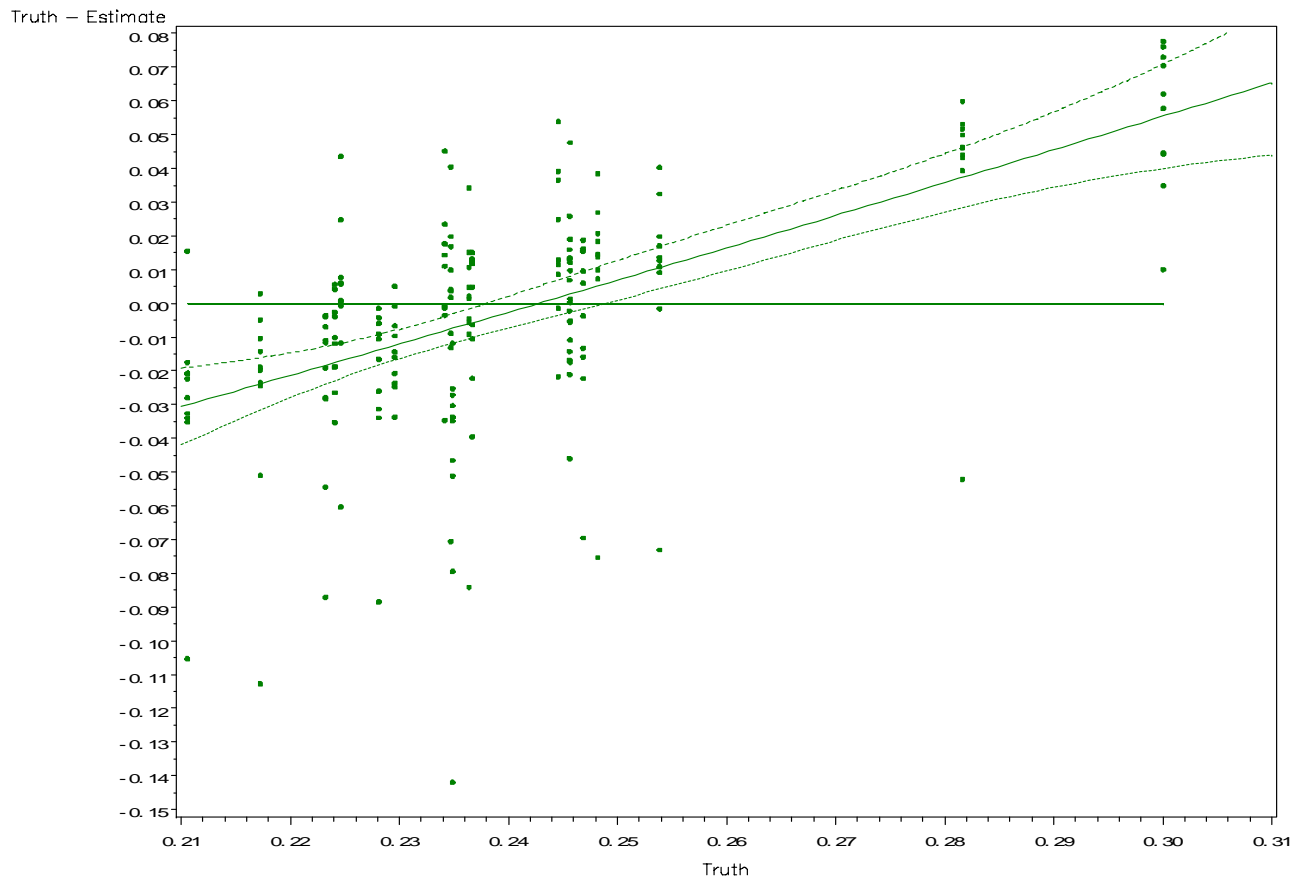
Difference between truth and estimated value of welfare estimate: FGT2
Ten different survey-based estimates for each of 20 target populations



Average difference: -0.0015

Figure 4: Checking for Bias

Difference between truth and estimated value of welfare estimate: GEO
Ten different survey-based estimates for each of 20 target populations



Average difference: -0.0024

Figure 5: Model Specification and Bias

HEADCOUNT: Difference in bias due to failure to account for unobserved location effects
Preferred model against model with no location variables

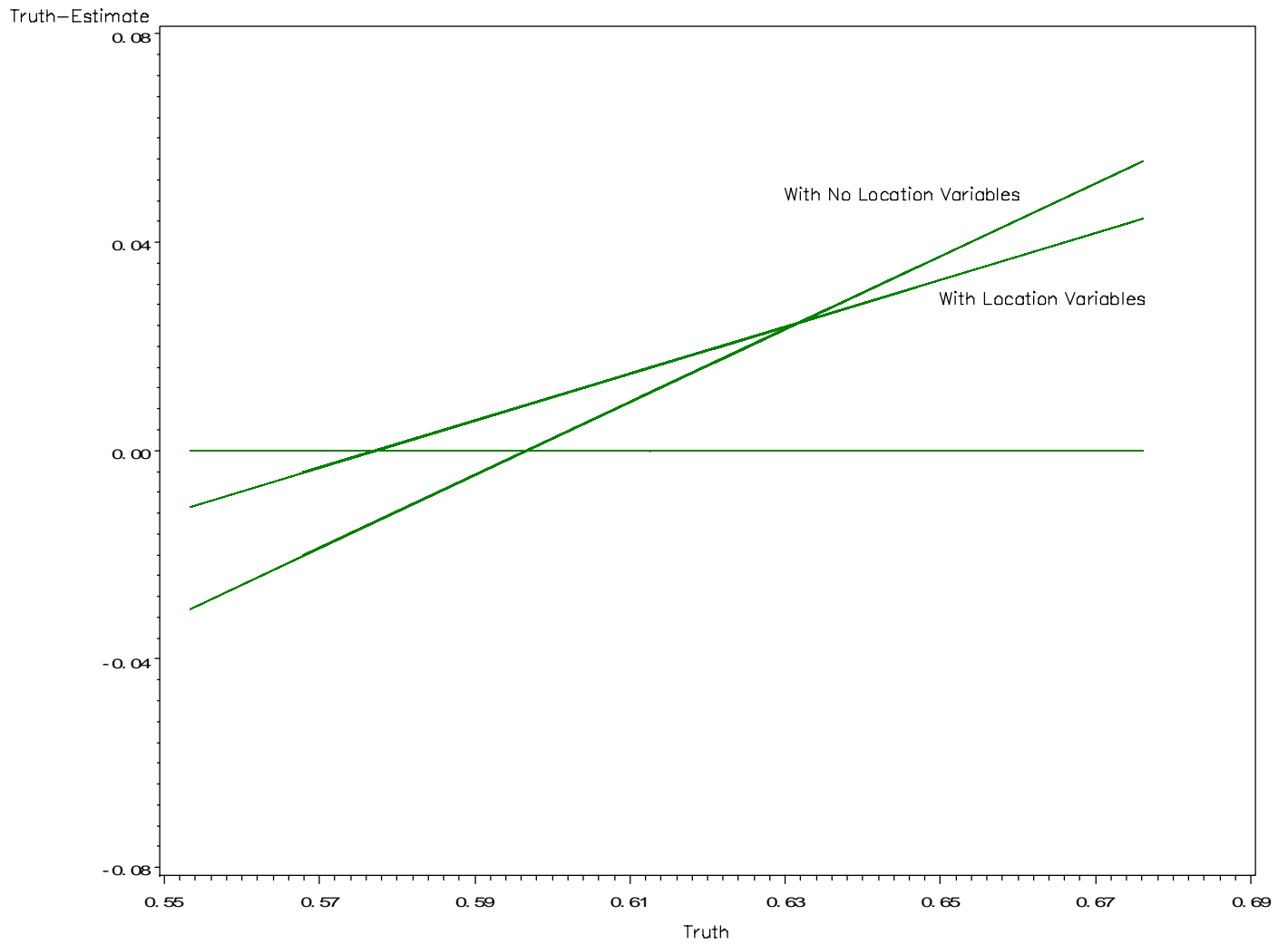


Table 6: Correlations Between Estimated and True Welfare Across Target Populations

Survey	Mean		Headcount		FGT2		GE0	
	Pearson	Spearman	Pearson	Spearman	Pearson	Spearman	Pearson	Spearman
1	0.58	0.53	0.64	0.58	0.73	0.75	0.14	-0.05
2	0.27	0.32	0.20	0.22	0.47	0.55	0.02	-0.01
3	0.68	0.69	0.62	0.61	0.54	0.45	0.03	0.14
4	0.50	0.54	0.59	0.57	0.33	0.29	-0.11	-0.06
5	0.67	0.67	0.75	0.69	0.71	0.67	-0.02	0.12
6	0.45	0.50	0.67	0.73	0.80	0.78	0.06	0.15
7	0.37	0.36	0.35	0.30	0.21	0.20	0.24	0.07
8	0.66	0.67	0.59	0.50	0.53	0.51	0.18	0.15
9	0.22	0.11	0.23	0.12	0.15	0.04	0.11	0.18
10	0.28	0.21	0.38	0.28	0.18	0.08	-0.17	-0.18
Average	0.47	0.46	0.50	0.46	0.46	0.43	0.05	0.05

Appendix 1

A Non-parametric Simulation Procedure

In this appendix we describe the procedure used for generating the welfare predictions reported in the paper. The procedure was developed to diminish the role of distributional assumptions and increase the role of bootstrapping.

A key aspect of the prediction is the way in which 'model error' is handled, or the inevitable deviation between estimated and true parameters.²⁰ So far we have accounted for model error using the estimated covariance matrices for the model parameters. Alternatively, sampling error of the parameter estimates can be simulated directly, by re-sampling the survey and re-estimation of the parameters, which is what we do in the current paper. The survey is resampled by parametric bootstrapping of the error term, based on an initial set of point estimates and residuals. This procedure also allows us to detect bias in the estimators for the parameters of the error model.

Starting from any given 'fake survey' the steps are as follows²¹:

1. For the current application, model selection must necessarily be a semi-automatic procedure. Thus we carry out an OLS regression of log per capita consumption on an extensive set of candidate variables.
2. Next we limit the number of covariates using a procedure for step-wise selection of regressors.
3. With the resulting set of regressors, we specify and estimate a linear mixed effect model accounting for both cluster random effects and household-level heteroskedasticity.²² We have used the following specification for heteroskedasticity:

$$\sigma_h = \alpha_0 e^{\alpha_1 \hat{y}_h}$$

where \hat{y}_h denotes the point estimate of household h 's log per capita consumption (pcx).

4. The estimation yields
 - point estimates for the regression coefficients, $\hat{\beta}$.
 - point estimates for log per capita expenditure, \hat{y} .
 - point estimates for the heteroskedasticity model, $\hat{\alpha}$.
 - the $\hat{\alpha}$ allows us to derive point estimates for the standard deviation of household-level errors, $\hat{\sigma}_s$.

²⁰ 'True' is interpreted here as the parameter estimates that would result from a sample consisting of the full population.

²¹ The computations have been carried out using R version 2.2.1 and the nlme package, version number 3.1.66. Script files of the procedure can be obtained upon request from the authors.

²² See Venables and Ripley (1997) and Bates and Pinheiro (1998). The procedures for estimating linear mixed effect models in R's nlme package can handle cluster random effects and household-level heteroskedasticity of a simple type.

- residuals, which we split into mean residuals per cluster, $\hat{\eta}$, the standard deviation of these, σ_{η} , and deviations from the cluster mean, $\hat{\varepsilon}$.
- the standardized household residuals, $\hat{\sigma}_{\varepsilon} = \frac{\hat{\varepsilon}}{\hat{\sigma}_s}$.

These estimates are used to check for bias in the estimation procedure. There is reason to expect such a bias, especially for the heteroskedasticity model and the variance of the cluster effects η .²³

5. The general idea to generate 100 samples by parametric bootstrapping using the above parameters as the 'true' model. We resample η 's from $\hat{\eta}$, standardized household residuals from $\hat{\sigma}_{\varepsilon}$, multiplying the latter with each households specific standard deviation from $\hat{\sigma}_s$. The total residual is added to \hat{y} to yield a new value for log per capita expenditure for each household. The new value is compatible with the model estimated under 3 above, and with the value of household regressors.
6. Each bootstrapped sample is used to re-estimate the model and the mean of the estimates is used to check for estimation bias. It turns out that the bias (if any) is small and inconsequential. Nevertheless, we have compensated for bias in the estimators for σ_{η} and α using the average bias found in this first round of simulations.

With the adjusted values for the variance estimators we again generate 100 samples by parametric bootstrapping.

7. For each sample we reestimate the model, resulting in point estimates for β , α , η , and σ_{ε} . These are used to impute log per capita consumption values for households in the 'census'. For census 'EAs' an η is drawn from the estimation result, for households a σ_{ε} is drawn and multiplied with the household-specific variance, using the current value of α . The sum of cluster and household 'error' is added to the systematic part of log per capita expenditure, based on the household regressors and the current value of β .

Thus we generate values of log per capita expenditure for all households in the census. Using these we compute welfare statistics (poverty and inequality measures). The tables and figures in the text represent means and standard deviations of the simulated welfare statistics thus generated.

²³ See Pfeifferman and Glickman(2004), and Rao (2003). The estimators for the regression coefficients are unbiased regardless of the error structure imposed.

Appendix References

- Bates, D.M. and Pinheiro, J.C. (1998) "Computational methods for multilevel models" available in PostScript or PDF formats at <http://franz.stat.wisc.edu/pub/NLME/>
- Pfeffermann, D., and Tiller, R. (2005). Bootstrap Approximation to Prediction MSE for State-Space Models with Estimated Parameters. *Journal of Time Series Analysis*, 26, 893-216.
- Pfeffermann, D., and Glickman, H. (2004). "Mean Square Error Approximation in Small Area Estimation By Use of Parametric and Nonparametric Bootstrap". Invited lecture at the Joint Statistical Meeting, Toronto.
- Rao, J.N.K. (2003) *Small Area Estimation*. Wiley: New York.
- Venables, W.N. and Ripley, B.D. (1997) *Modern Applied Statistics with S-plus*. 3rd Edition, Springer-Verlag.

Appendix 2: OLS Regression Results of Consumption Models

Table 1: Pseudo Survey 1

Dependent Variable: Log Per Capita Expenditure

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	5.998746	0.376066	15.951	< 2e-16	***
hsize	-0.088087	0.013425	-6.562	1.37e-10	***
onlyindhead	-0.357614	0.187783	-1.904	0.057450	.
refrig	0.164402	0.076970	2.136	0.033187	*
toilet	-0.096050	0.052603	-1.826	0.068475	.
vehicle	0.203101	0.088630	2.292	0.022359	*
bilinghead	-0.341641	0.080568	-4.240	2.67e-05	***
rethead	0.092900	0.059246	1.568	0.117526	
av_femhead	-0.898957	0.371149	-2.422	0.015798	*
av_onlyindhead	2.250072	0.566152	3.974	8.13e-05	***
av_primedhead	0.774239	0.260069	2.977	0.003056	**
av_rethead	0.786780	0.223840	3.515	0.000481	***
av_runwater	-0.098425	0.066368	-1.483	0.138717	
rhsiz2	0.796609	0.167641	4.752	2.66e-06	***
rroompp	-0.174065	0.039715	-4.383	1.44e-05	***
rroompp2	0.011750	0.003473	3.384	0.000773	***

Multiple R-Squared: 0.4838, Adjusted R-squared: 0.4678

Table 2: Pseudo Survey 2

Dependent Variable: Log Per Capita Expenditure

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	9.129474	0.804244	11.352	< 2e-16	***
hsize	-0.096499	0.014865	-6.492	2.12e-10	***
gasstove	0.172803	0.070264	2.459	0.014270	*
refrig	0.133641	0.081375	1.642	0.101186	
toilet	0.087655	0.059192	1.481	0.139298	
adultfracf	0.327968	0.159454	2.057	0.040243	*
av_adultfracm	0.747587	0.468641	1.595	0.111320	
av_agehead	-0.033981	0.007541	-4.506	8.29e-06	***
av_concreterooft	-0.382385	0.207337	-1.844	0.065759	.
av_femhead	-2.605026	0.637204	-4.088	5.09e-05	***
av_primedhead	-0.659667	0.308155	-2.141	0.032800	*
av_radio	-0.874114	0.318263	-2.747	0.006249	**
av_rethead	0.451829	0.286645	1.576	0.115622	
av_runwater	-0.179179	0.086103	-2.081	0.037964	*
av_television	0.776940	0.212897	3.649	0.000292	***
av_waterheater	1.502314	0.854344	1.758	0.079308	.
rhsiz2	0.953988	0.147147	6.483	2.23e-10	***
rroompp	-0.027115	0.017004	-1.595	0.111454	
ragehead2	123.342227	50.256095	2.454	0.014470	*

Multiple R-Squared: 0.4788, Adjusted R-squared: 0.4593

Table 3: Pseudo Survey 3

Dependent Variable: Log Per Capita Expenditure

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	7.740004	0.553045	13.995	< 2e-16	***
hsize	-0.111892	0.012125	-9.228	< 2e-16	***
blender	0.142074	0.069276	2.051	0.040833	*
brickwall	-0.123116	0.065063	-1.892	0.059067	.
gasstove	0.231063	0.072605	3.182	0.001556	**
naturalroof	-0.169465	0.071431	-2.372	0.018070	*
onlyindhead	0.242028	0.166921	1.450	0.147733	
radio	0.140417	0.055806	2.516	0.012193	*
stereo	0.247070	0.116874	2.114	0.035038	*
adultfracf	0.302865	0.165445	1.831	0.067787	.
bilinghead	0.163705	0.073534	2.226	0.026468	*
agehead	-0.002257	0.001585	-1.424	0.155226	
secedhead	0.227859	0.118303	1.926	0.054693	.
av_agehead	-0.015256	0.006668	-2.288	0.022575	*
av_blender	-1.091239	0.259010	-4.213	3.02e-05	***
av_concreteroof	1.030535	0.205624	5.012	7.63e-07	***
av_femhead	-0.657499	0.421795	-1.559	0.119708	
av_hsize	-0.096361	0.038338	-2.513	0.012285	*
av_onlyindhead	-0.539298	0.359583	-1.500	0.134336	
av_primedhead	-0.386760	0.255997	-1.511	0.131505	
av_radio	-0.745915	0.219001	-3.406	0.000715	***
av_refrig	0.870410	0.258107	3.372	0.000807	***
av_television	0.807982	0.192275	4.202	3.16e-05	***
av_toilet	-0.258594	0.096860	-2.670	0.007851	**
av_waterheater	-1.194664	0.657062	-1.818	0.069666	.
rhsiz2	0.949978	0.146055	6.504	1.99e-10	***

Multiple R-Squared: 0.5511, Adjusted R-squared: 0.5274

Table 4: Pseudo Survey 4

Dependent Variable: Log Per Capita Expenditure

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	5.061854	0.351890	14.385	< 2e-16	***
hsize	-0.109834	0.016429	-6.685	6.35e-11	***
refrig	0.174286	0.076323	2.284	0.022831	*
toilet	0.161254	0.054947	2.935	0.003497	**
adultfracm	0.320246	0.139893	2.289	0.022495	*
adultfracf	0.293536	0.138096	2.126	0.034042	*
bilinghead	0.143261	0.062064	2.308	0.021403	*
secedhead	0.205535	0.105298	1.952	0.051520	.
av_agehead	0.014903	0.007363	2.024	0.043521	*
av_blender	0.423415	0.159784	2.650	0.008314	**
av_brickwall	0.382044	0.128597	2.971	0.003117	**
av_radio	-0.727830	0.218147	-3.336	0.000914	***
rhsiz2	0.476885	0.148333	3.215	0.001392	**
rroompp	-0.140513	0.045565	-3.084	0.002161	**
rroompp2	0.012268	0.004478	2.740	0.006379	**

Multiple R-Squared: 0.4315, Adjusted R-squared: 0.4151

Table 5: Pseudo Survey 5

Dependent Variable: Log Per Capita Expenditure

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	6.20858	0.33501	18.533	< 2e-16	***
hsize	-0.10914	0.01331	-8.198	2.22e-15	***
blender	0.17330	0.06220	2.786	0.00554	**
brickwall	0.19870	0.06127	3.243	0.00126	**
onlyindhead	-0.31920	0.16104	-1.982	0.04804	*
toilet	0.09907	0.05699	1.738	0.08279	.
adultfracm	0.26519	0.13636	1.945	0.05239	.
av_adultfracm	1.05350	0.38360	2.746	0.00625	**
av_blender	-0.36338	0.16296	-2.230	0.02621	*
av_femhead	-0.88381	0.36526	-2.420	0.01590	*
av_refrig	1.56893	0.30584	5.130	4.21e-07	***
av_runwater	0.19768	0.07834	2.524	0.01194	*
av_secedhead	-0.88101	0.49439	-1.782	0.07538	.
av_toilet	-0.38558	0.11117	-3.468	0.00057	***
av_washmachine	-1.43055	0.49677	-2.880	0.00416	**
rhsiz2	0.72648	0.15162	4.791	2.21e-06	***
rroompp	-0.04117	0.01615	-2.550	0.01109	*

Multiple R-Squared: 0.5331, Adjusted R-squared: 0.5176

Table 6: Pseudo Survey 6

Dependent Variable: Log Per Capita Expenditure

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	4.830e+00	4.540e-01	10.639	< 2e-16	***
hsize	-1.031e-01	1.365e-02	-7.555	2.13e-13	***
blender	1.693e-01	6.489e-02	2.608	0.009384	**
onlyindhead	-3.751e-01	1.941e-01	-1.933	0.053881	.
refrig	1.485e-01	7.901e-02	1.879	0.060809	.
bilinghead	-3.069e-01	7.068e-02	-4.342	1.73e-05	***
agehead	-6.775e-03	2.913e-03	-2.325	0.020469	*
av_adultfracm	2.464e+00	6.953e-01	3.545	0.000432	***
av_agehead	-1.184e-02	5.958e-03	-1.987	0.047493	*
av_blender	-5.047e-01	1.965e-01	-2.569	0.010503	*
av_brickwall	1.187e+00	2.092e-01	5.671	2.45e-08	***
av_concreterooft	-7.636e-01	2.516e-01	-3.035	0.002537	**
av_onlyindhead	3.661e+00	5.887e-01	6.219	1.09e-09	***
av_rechead	1.371e+00	2.384e-01	5.752	1.57e-08	***
av_refrig	4.606e-01	3.069e-01	1.501	0.134090	.
av_washmachine	-7.053e-01	3.694e-01	-1.909	0.056798	.
av_waterheater	2.058e+00	7.781e-01	2.645	0.008436	**
rhsiz2	6.923e-01	1.459e-01	4.746	2.74e-06	***
rroompp	-4.672e-02	1.594e-02	-2.931	0.003541	**
ragehead2	-1.285e+02	8.692e+01	-1.479	0.139890	.

Multiple R-Squared: 0.4965, Adjusted R-squared: 0.4766

Table 7: Pseudo Survey 7

Dependent Variable: Log Per Capita Expenditure

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	7.05900	0.43468	16.240	< 2e-16	***
hsize	-0.12896	0.01467	-8.793	< 2e-16	***
brickwall	0.12956	0.06725	1.927	0.054605	.
refrig	0.27110	0.08101	3.347	0.000882	***
toilet	0.10500	0.06705	1.566	0.117989	.
rechead	0.11263	0.05835	1.930	0.054186	.
av_brickwall	0.44800	0.19190	2.335	0.019975	*
av_concreterooft	-0.65035	0.24228	-2.684	0.007518	**
av_femhead	-2.13496	0.43132	-4.950	1.03e-06	***
av_hsize	0.16780	0.04718	3.556	0.000414	***
av_primedhead	0.73362	0.31380	2.338	0.019801	*
av_radio	-0.41700	0.19357	-2.154	0.031714	*
av_secedhead	1.06547	0.75789	1.406	0.160414	.
av_secplusedhead	-2.31016	1.26275	-1.829	0.067947	.
av_toilet	-0.33099	0.12981	-2.550	0.011084	*
av_waterheater	-1.91772	0.77601	-2.471	0.013809	*
rhsiz2	0.51461	0.14845	3.467	0.000574	***
rroompp	-0.04888	0.01765	-2.769	0.005839	**

Multiple R-Squared: 0.4735, Adjusted R-squared: 0.4549

Table 8: Pseudo Survey 8

Dependent Variable: Log Per Capita Expenditure

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	6.734e+00	5.829e-01	11.552	< 2e-16	***
hsize	-1.208e-01	1.689e-02	-7.151	3.22e-12	***
radio	1.080e-01	6.370e-02	1.695	0.09076	.
refrig	2.748e-01	8.626e-02	3.186	0.00154	**
toilet	1.568e-01	7.117e-02	2.203	0.02806	*
vehicle	2.872e-01	1.095e-01	2.623	0.00898	**
agehead	-7.636e-03	3.424e-03	-2.230	0.02619	*
av_adultfracm	1.856e+00	9.437e-01	1.967	0.04976	*
av_concreterooft	8.002e-01	1.790e-01	4.472	9.70e-06	***
av_femhead	-1.495e+00	5.201e-01	-2.875	0.00422	**
av_primedhead	-1.095e+00	4.020e-01	-2.724	0.00668	**
av_rehead	5.684e-01	2.779e-01	2.045	0.04139	*
av_runwater	-1.586e-01	8.212e-02	-1.931	0.05410	.
av_secedhead	2.328e+00	7.829e-01	2.974	0.00309	**
av_toilet	-2.154e-01	1.340e-01	-1.608	0.10844	.
rhsiz2	8.414e-01	1.902e-01	4.424	1.20e-05	***
rroompp	-7.209e-02	3.950e-02	-1.825	0.06864	.
rroompp2	6.130e-03	3.114e-03	1.968	0.04962	*
ragehead2	-1.873e+02	1.038e+02	-1.805	0.07173	.

Multiple R-Squared: 0.4414, Adjusted R-squared: 0.4205

Table 9: Pseudo Survey 9

Dependent Variable: Log Per Capita Expenditure

Coefficients:

	Value	Std.Error	DF	t-value	p-value
(Intercept)	5.086357	0.1885547	441	26.975497	0.0000
hsize	-0.141745	0.0124185	441	-11.414072	0.0000
brickwall	0.104505	0.0600241	441	1.741055	0.0824
gasstove	0.135917	0.0672063	441	2.022382	0.0437
onlyindhead	-0.895540	0.1898896	441	-4.716112	0.0000
radio	0.137231	0.0543460	441	2.525141	0.0119
adultfracf	0.402884	0.1555154	441	2.590636	0.0099
bilinghead	-0.111148	0.0660533	441	-1.682702	0.0931
secedhead	0.260845	0.1083821	441	2.406719	0.0165
av_hsize	0.098639	0.0312940	44	3.152021	0.0029
av_runwater	-0.149705	0.0722344	44	-2.072487	0.0441
av_secedhead	1.286449	0.3965916	44	3.243761	0.0023
av_television	-0.318822	0.1260081	44	-2.530169	0.0151
av_washmachine	1.140216	0.2628267	44	4.338280	0.0001
rhsiz2	0.653687	0.1320114	441	4.951745	0.0000

Multiple R-Squared: 0.506, Adjusted R-squared: 0.491

Table 10: Pseudo Survey 10

Dependent Variable: Log Per Capita Expenditure

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	6.31149	0.23697	26.634	< 2e-16	***
hsize	-0.11552	0.01597	-7.232	1.87e-12	***
naturalroof	-0.18068	0.07998	-2.259	0.024322	*
television	0.14613	0.05767	2.534	0.011593	*
vehicle	0.26146	0.09715	2.691	0.007363	**
bilinghead	-0.15631	0.07185	-2.175	0.030083	*
av_adultfracm	-1.67378	0.69667	-2.403	0.016655	*
av_blender	-0.65744	0.19602	-3.354	0.000859	***
av_brickwall	0.22799	0.12851	1.774	0.076677	.
av_radio	-0.59248	0.21000	-2.821	0.004978	**
av_roompp	0.64006	0.23260	2.752	0.006150	**
av_secedhead	1.37118	0.53967	2.541	0.011371	*
rhsiz2	0.72202	0.17679	4.084	5.17e-05	***
rroompp	-0.03031	0.01717	-1.765	0.078153	.

Multiple R-Squared: 0.4344, Adjusted R-squared: 0.4193